OXFORD

Briefings in Bioinformatics, 20(4), 2019, 1384–1394

doi: 10.1093/bib/bby007
Advance Access Publication Date: 31 January 2018
Paper

How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives

Alessandra Dal Molin and Barbara Di Camillo

Corresponding author. Barbara Di Camillo, Department of Information Engineering, University of Padova, via Giovanni Gradenigo 6, Padova, Italy. Tel: +39 049 8277671; E-mail: barbara.dicamillo@unipd.it

Abstract

The sequencing of the transcriptome of single cells, or single-cell RNA-sequencing, has now become the dominant technology for the identification of novel cell types in heterogeneous cell populations or for the study of stochastic gene expression. In recent years, various experimental methods and computational tools for analysing single-cell RNA-sequencing data have been proposed. However, most of them are tailored to different experimental designs or biological questions, and in many cases, their performance has not been benchmarked yet, thus increasing the difficulty for a researcher to choose the optimal single-cell transcriptome sequencing (scRNA-seq) experiment and analysis workflow. In this review, we aim to provide an overview of the current available experimental and computational methods developed to handle single-cell RNA-sequencing data and, based on their peculiarities, we suggest possible analysis frameworks depending on specific experimental designs. Together, we propose an evaluation of challenges and open questions and future perspectives in the field. In particular, we go through the different steps of scRNA-seq experimental protocols such as cell isolation, messenger RNA capture, reverse transcription, amplification and use of quantitative standards such as spike-ins and Unique Molecular Identifiers (UMIs). We then analyse the current methodological challenges related to preprocessing, alignment, quantification, normalization, batch effect correction and methods to control for confounding effects.

Key words: single-cell RNA-seq; experimental protocols; experimental design; normalization; spike-ins; Unique Molecular Identifiers

Introduction

Single-cell transcriptome sequencing (scRNA-seq) has recently emerged as a powerful tool to study individual cell transcriptomes on a large scale. The high resolution of this technology offers, beyond the advantages of bulk RNA-seq experiments, the unique opportunity to dissect the transcriptional landscape of single cells and opens the door to the study of novel biological questions. Indeed, scRNA-seq enables the study of gene expression dynamics and differentiation trajectories, the investigation of regulatory processes and transcriptional kinetics as well as the identification of novel cell types. In fact, while in bulk experiments gene expression is averaged across a cell population, with scRNA-seq it is possible to look at the real 'state' of single cells. Indeed, tissues and organs are complex multicell systems made of multiple subpopulations of cells exhibiting different system states, i.e. expression profiles, which may vary from cell to cell because of the presence of heterogeneous subpopulations, cells at different stages of the cell cycle or cells responding to different stimuli. Moreover, cells are spatially and temporally organized and able to communicate and interact with each other to orchestrate self-assembly and response to stimuli as a whole. Therefore, studying gene expression at single-cell level and comparing gene expression profiles

Alessandra Dal Molin is a Postdoctoral Research Fellow at the Information Engineering Department of the University of Padova (Padova, Italy). She is working on bulk RNA-seq and single-cell RNA-seq snapshot and time series data analysis, benchmarking existing methods and developing new computational tools.

Barbara Di Camillo is an Associate Professor at the Information Engineering Department of the University of Padova (Padova, Italy). She is working on machine learning and modelling of biological systems and development of statistical methods for high-throughput data analysis specifically tailored for bulk and single-cell RNA-seq data.

Submitted: 31 October 2017; **Received (in revised form):** 27 December 2017

between individual cells allow discovering previously undetected populations and unveiling new regulatory paths.

The interest in the study of gene expression in single cells began already in the seventies, but only recently, it has gained a widespread popularity, with an exponential growth of published studies in fields such as neurobiology and oncology [1–3] and a *plethora* of published protocols and software tools. At present, most protocols and methodologies are tailored to specific experimental designs or biological questions, and in many cases, their performance has not been benchmarked yet, thus increasing the difficulty for a researcher to choose the optimal scRNA-seq experiment and analysis workflow.

Despite its power and high resolution, scRNA-seq has some open challenges related to the higher level of technical noise and data complexity with respect to bulk RNA-seq, related for instance to 'dropout' events, i.e. the inability to capture some transcript messenger RNAs (mRNAs), especially if at low abundance, mainly because of reverse transcription (RT) inefficiency and low sequencing depth. In this review, we provide an overview of the experimental and computational methods alongside their strengths and weaknesses, suggesting different analysis frameworks depending on the specific biological questions, and evaluating challenges and open questions together with the future perspectives in the field.

scRNA-seq experimental protocols

A typical scRNA-seq experimental workflow begins with the dissociation of cells from a tissue and the isolation of single cells with specific devices; then, mRNAs are captured for RT and amplification and, finally, the synthesized complementary DNA (cDNA) molecules undergo library preparation for sequencing.

Cell isolation

The first step of scRNA-seq is cell isolation, and its efficiency depends on the protocol being used (Table 1).

Early methods for single-cell isolation include micropipetting, micromanipulation and laser capture microdissection [4–6]. These methods are low throughput, technically challenging and entail a laborious time-consuming procedure with respect to more recently developed methods but are still used when the number of cells to analyse is low (e.g. rare cells) [4–6].

Fluorescence-activated cell sorting (FACS) is a specialized type of flow cytometry that provides a method for sorting heterogeneous mixtures of cells, one cell at a time, based on cell size and fluorescence, at higher throughput and more quickly (in the order of thousands or tens of thousands cells/day) with respect to its predecessors. Usually—but not exclusively—fluorescently labelled antibodies are used to isolate cells, allowing the measurement of labelled protein fluorescence and even the identification of novel subsets of cells in well-characterized populations; whereas index sorting enables the isolation of cells based on a particular transcriptional state (e.g. cell cycle state) or morphology [5]. Potential limitations of FACS include the need for specific antibodies and the possible interference of these antibodies with downstream analyses, but also the large volume of input material required (microlitre or even millilitre order), which hampers the isolation of cells deriving from extremely low-volume samples or the isolation of rare cells [5].

More recently, microfluidic devices have been introduced and are rapidly taking place as the preferable technique to isolate cells, as they require smaller volumes of reagents with respect to FACS and other previously used methods. In microfluidic devices, a hydrodynamic flux permits isolation and processing of single cells in channels of dimension of tens to hundreds of microns, thus comparable with the size of a single cell. Moreover, microfluidic devices can also automate some downstream RNA processing reactions for sequencing (see section 'Use of quantitative standards') and allow measuring and controlling extracellular reagents concentration [7].

Table 1 shows the main microfluidic systems distinguished by capturing strategy, the throughput and the speed of the process, the possibility to barcode the cells and to process one or more experimental conditions in one run. There are two main kinds of capturing strategies: the first uses array-based techniques where cell isolation is performed by capturing cells within tailored microfluidic array chambers and the second uses droplet-based techniques, where each cell is encapsulated into a microfluidic droplet. Cell barcoding allows individual labelling of each cell sample, thus enabling cell pooling and multiplexing within sequencing.

The most popular microfluidic system is the C1™ System by Fluidigm®, which allows to process up to 800 single cells through an array-based strategy, with the possibility to load up to two experimental conditions in one run. This system is the most versatile one, as it allows using different types of protocols for scRNA-seq in addition to the one supplied by the company, but, because of the characteristics of the array, it is suitable for spherical cells of specific size from a minimum of 5 μm to a maximum of 25 μm. The cost of the machine is approximately >150 000 dollars [8].

Another array-based microfluidic system of more recent release is the ICELL8™ (Wafergen Biosystems®). This system further increases the throughput with respect to the C1 System, with the ability of isolating thousands of cells in few hours and of including downstream sample preparation steps for sequencing. In addition, it has the capability of loading up to eight experimental conditions across one array through a single needle, called MultiSample NanoDispenser (MSND). It is not clear if

Table 1. Main features of most popular single-cell isolation microfluidic systems

System/company	Strategy	Cell barcoding	#Cells captured per run	Time per run (hh:mm)	#Different samples per run	Indicative instrument price ($)
C1 System/Fluidigm	Array-based	No	96	01:00	1	∼150–200k
		Yes	800	24:00	2	
ICELL8/Wafergen	Array-based	Yes	5184	07:00	8	∼250k
DEPArray/MENARINI Silicon Biosystem	Array-based	No	96	02:30	1	∼100k
Single cell RNA-Seq System/Dolomite Bio	Droplet-based	Yes	∼15 000	00:15	Multiple	∼50k
Chromium/10X Genomics	Droplet-based	Yes	100–48 000	00:10	8	∼125k
ddSEQ/Bio-Rad+Illumina	Droplet-based	Yes	96	<00:05	4	∼60k

there are potential carry-over contaminations of the dispenser because of the consecutive loading of different samples. A potential criticality of this system is that the final library is created by pooling all array wells (corresponding to all the different experimental conditions) in a single tube for sequencing, leading to the possibility that low abundant transcripts may not be represented in the sequencing reads. This issue will be discussed also later in the manuscript. The price of the instrument is >200 000 dollars, but the cost of the experiment is reduced by the creation of a single sequencing library.

The DEParray™ System (Menarini-Silicon Biosystems) technology is based on the ability of a non-uniform electric field to exert forces on suspended cells, following an electrokinetic principle called dielectrophoresis. The creation of this electric field allows the trapping, manipulation (individual cells of interest may be moved to specific locations) and recovery of individual cells. Among the described array-based systems, this is the cheapest one, even if the number of cells analysed per run is not high but may be a good option if the aim of the study is, e.g., the study of cell–cell interactions or the study of rare, specific cells. In particular, for the latter purpose, Menarini-Silicon Biosystems has recently acquired the CELLSEARCH® Circulating Tumor Cells (CTC) System, the only FDA-cleared platform for CTC testing.

Among the droplet-based systems, the single-cell RNA-seq system by Dolomite Bio implements the Drop-seq protocol [9] with the novelty of reusable and higher throughput (4000 droplets/s) glass chips. Cell barcoding is possible, whereas RT and amplification reactions are supposed to be done manually after cells are isolated in droplets, with other commercial or in-house protocols. Recently, this system has been improved by including the capability to perform high-throughput isolation of single nuclei with DroNc-Seq [10].There could be potential criticalities because of inaccurate washing of the glass chips or because of the use of non-standard post-droplet in-house protocols. Other droplet-based systems, the ddSEQ™ (by Illumina® and Bio-Rad®) and the Chromium™ (10X Genomics®), potentially solve this issue, as they provide specific kits for mRNA processing after cell isolation. In general, the advantages of droplet-based systems are that the price for the instruments and protocols is lower than their array-based counterparts, and the number of analysed cells per run may be high; moreover, they can process multiple different conditions in one run. On the other hand, as for the ICELL8 array-based system, the final pooling of all droplets and the coverage bias of related protocols may lead to the under-representation of low abundant transcripts in the final library, thus hampering of the investigation of the complete transcriptional landscape of the single cells.

mRNA capture, RT and amplification

Together with the isolation of single cells, performed using one of the previously described strategies, mature mRNAs must be captured, reverse transcribed into cDNAs and amplified.

Table 2 shows the most popular scRNA-seq protocols based on different strategies for RT, cDNA synthesis and amplification, together with the possibility to accommodate sequence-specific barcodes (detailed later in section 'Transcripts quantification') or the ability to process pooled samples.

Currently, most of the single-cell microfluidic devices reported in Table 1 supply their own scRNA-seq protocols whose features recall the ones reported for protocols of Table 2.

Many devices make use of specific barcodes, which allow the capture of multiple cells and mRNAs simultaneously, procedure which is called 'multiplexing'. Islam *et al.* [18] published the first scRNA-seq protocol able to perform multiplexing using a unique template switching oligo in each well of a 96-well array [19]. Later, both the inDrop and Drop-seq protocols also included barcoded cDNA preparation within the droplets, while, in a recent alternative approach, cells are deposited into picolitre wells that contain barcoded beads and reagents [15]. Another recent strategy is the use of combinatorial *in situ* barcoding, adopted in the single-cell combinatorial indexing RNA-sequencing [20] and SPLiT-Seq [16] methods. Here, single cells follow a number of barcoding rounds (random split in mini-pools and distribution in multiwell arrays with unique barcodes), so that at the end of the procedure, they are uniquely labelled [19].

Usually, RT of mRNAs is performed using an oligo-dT primer. This is done to avoid the capture of other structural RNAs such as ribosomal RNAs and transfer RNAs, which account for the majority of cellular RNAs. However, using this strategy, protein-coding RNAs without the poly-A tail are not captured [21]. Moreover, the use of oligo-dT primers suffers of low capture efficiency, which, for current protocols, is reported to be around 10–15% [22]. Indeed, while bulk RNA-seq studies usually average the transcriptomes of millions of cells, in scRNA-seq, the aim is to capture the transcriptome of a single cell, where the majority of mRNAs are present only in few copies (less than five transcripts) [5]. Klein *et al.* [23] and Macosko *et al.* [9] estimated that their droplet-based methods inDrop and Drop-seq are able to capture about 7.1 and 12.8% of the mRNAs in the cell, respectively, but even the other recently developed scRNA-seq protocols suffer of low mRNA capture efficiency [22, 24, 25].

In scRNA-seq experiments, as well as in bulk RNA-seq, the step of RT to cDNA is necessary, because of RNA instability, which makes it difficult to use RNA-dependent polymerases. Then, cDNAs have to be amplified to obtain the needed quantities for sequencing. The main strategies adopted for cDNA

Table 2. Main features of most popular single-cell protocols for mRNA capture, RT and amplification (latest update: December 2017)

Method	Reference	RT primers	cDNA synthesis	Amplification method	UMIs	Transcript coverage	Sample pooling
Tang *et al.*	Tang *et al.* [11]	Poly(T) + poly(A)	Poly(A) tailing	PCR	No	Nearly full-length	No
Quartz-seq2	Sasagawa *et al.* [12]	Poly(T) + poly(A)	Poly(A) tailing	PCR	Yes	3′-end	Yes
STRT-seq	Islam *et al.* [18]	Poly(T)	Template switching	PCR	Yes	5′-end	Yes
SMART-seq2	Picelli *et al.* [13]	Poly(T)	Template switching	PCR	No	Full-length	No
SCRB-seq	Soumillon *et al.* [14]	Poly(T)	Template switching	PCR	Yes	3′-end	No
Drop-seq	Macosko *et al.* [9]	Oligo-dT	Template switching	PCR	Yes	3′-end	Yes
Seq-Well	Gierahn *et al.* [15]	Poly(T)	Template switching	PCR	Yes	3′-end	Yes
SPLiT-seq	Rosenberg, Roco *et al.* [16]	Poly(T) (*in situ*)	Template switching	PCR	Yes	3′-end	Yes
CEL-seq2	Hashimshony *et al.* [17]	Poly(T)	IVT	IVT	Yes	3′-end	Yes

synthesis and amplification are the template switching coupled with polymerase chain reaction (PCR) and *in vitro* transcription (IVT) (Table 2).

The template-switching method exploits the two intrinsic properties of Moloney murine leukemia virus (MMLV). MMLV reverse transcriptase is able to add at the 5'-end of the RNA template, which corresponds to the 3'-end of the new cDNA strand, a few non-templated cytosines. These cytosines serve as an extended template for a helper oligonucleotide (called Template Switching Oligo - Locked Nucleic Acid) that allows the reverse transcriptase to 'switch' the template and synthesize the new cDNA strand [25]. This approach, used for example in the Smart-seq2 protocol [13], allows full-length amplification of the transcripts, differently from conventional approaches that often lead to premature termination of RT and exponential decrease in coverage towards one end of the transcripts. Although this method allows obtaining full-length transcripts, it does not take into account for 'failed template switching' cDNAs, which could instead contribute to the overall gene expression of the sample (Hebenstreit, 2012). Template switching is usually coupled with PCR to perform cDNA amplification. PCR, although being a fast and widely used method, has the disadvantage of leading to exponential amplification, bringing highly expressed transcripts to be over-represented in the final library [26]. Moreover, because of the different efficiency on different transcripts species, PCR amplification causes quantification biases, together with the loss of information of the original transcripts abundance and the accumulation of non-specific transcript fragments.

An alternative approach for cDNA synthesis and amplification, used in CEL-seq2 protocol [17], is the linear amplification through IVT. IVT consists in a first step of RT and cDNA synthesis using an Oligo(dT) primer, which contains a T7 promoter. The T7 polymerase repeatedly binds to the promoter and amplifies RNA, which eventually undergoes a final step of RT. The advantage of IVT lies in the linear amplification which, in contrast to PCR, does not exponentially deplete sequences that are difficult to process even if the protocol still has lower sensitivity for lowly expressed transcripts [25]. IVT is more labour intensive than PCR, as it requires an additional round of RT of the amplified RNA, often resulting in premature termination and thus to an accumulation of 3'enriched RNA fragments (strong 3'-bias), failing to detect the full transcriptome landscape of the sample [6, 21, 27].

Regarding strand specificity, both template switching and IVT theoretically maintain the strand information, but in most scRNA-seq protocols, fragmentation is performed only after transcript amplification (this to avoid further transcript loss), meaning that strand specificity is usually lost. The only way to preserve directional information is by selectively peeking either 5′ or 3′ ends after fragmentation, thus loosing, however, the full-length coverage [27].

Use of quantitative standards

The high technical variability introduced by the different processing steps hampers the ability to quantify transcript abundance accurately. Currently, possible solutions to these issues are the addition of quantitative standards like RNA spike-ins and/or Unique Molecular Identifiers (UMIs).

Spike-ins are artificial RNA molecules, which are added to cell lysate at known quantity and are subjected to all experimental steps after cell isolation. The purpose of using these molecules is to provide information about the relationship between the input number of molecules and the observed number of sequencing reads.

The most popular set of spike-ins is the set of 92 single-isoform synthetic RNAs of the External RNA Controls Consortium (ERCC) [28]. Spike-ins are supposed to act as endogenous RNAs but are not identical to them. Indeed, they have lengths from 250 to 2000 nucleotides, relatively short in comparison with mammalian genes and GC content between 30 and 50%; they have short poly-A tails (∼19–25 nucleotides, an order of magnitude less than endogenous RNA) and lack the 5' cap. Then, poly-T priming and template switching, which is required in several protocols, may be less efficient for them than for endogenous mRNAs, leading all experimental protocols to appear less sensitive than they are [29]. Another complication of using spike-ins is that they are typically added to the single-cell samples at high relative concentrations and, consequently, they take up a relatively large proportion of reads. Thus, not all protocols can accommodate their use, such as the recent droplet-based technologies [30]. Moreover, as the spike-ins are added to the cell lysis buffer, they cannot be used to estimate stochastic dropout of RNA molecules. The development of a set of spike-ins specifically designed for scRNA-seq could possibly overcome the above-described limitations.

Recently, Paul *et al.* [31] published a new set of spike-ins: 69 spike-in RNA variants named SIRVs, consisting of seven artificial genes with 6–18 transcript variants each, to mimic the transcription and splicing complexity. The GC content and length of the spike-ins and of their poly-A tails are similar to the ERCC spike-ins, thus bringing along the same practical limitations. In a recent study, the performance of ERCC spike-ins and SIRVs was compared, resulting in lower accuracy of SIRVs explained by ambiguous read mapping of alternative isoforms, especially when using 3'end-biased protocols [29].

More recently, researchers at the Garvan Institute of Medical Research have developed a set of spike-in RNA standards, termed 'sequins' (sequencing spike-ins), that represent full-length spliced mRNA isoforms [32]. Sequins have an entirely artificial sequence with no homology to natural reference genomes, but they align to gene loci encoded on an artificial *in silico* chromosome. The combination of multiple sequins across a range of concentrations emulates alternative splicing and differential gene expression. To the best of our knowledge, sequins have not been benchmarked yet.

Another type of quantitative standard used in scRNA-seq is the UMI. They are nucleotide sequences of length 4 up to 12 nucleotides, which are incorporated into the primer before RT, to uniquely barcode the 5' or the 3' end of each individual mRNA copy of each transcript. The underlying idea is to enable the quantification of each transcript based on the number of different UMIs, so to avoid the bias originated from PCR amplification. In a simple hypothetical scenario where transcript X is expressed with four copies and transcript Y with six copies, the two transcripts are barcoded with four and six different UMIs before amplification. As shown in Figure 1A, after the amplification step the information about real abundances of transcript copies would have been lost if UMIs had not been used. Instead, by using UMIs, the information about real abundances is theoretically preserved, and the counting of unique UMIs after sequencing results in a level of expression of four for transcript X and six for transcript Y.

To avoid underestimation of the original number of transcripts for highly expressed genes, the length n of the UMIs must be chosen, so that the number of unique barcodes (equal to 4^n, where $4 =$ number of azotate bases) is higher than the number of the transcript expressed at the highest level [22]. For example, UMIs of length $n \geq 7$ should be used, if the expected expression for the maximum expressed transcript is 10^4 copies (Figure 1B).

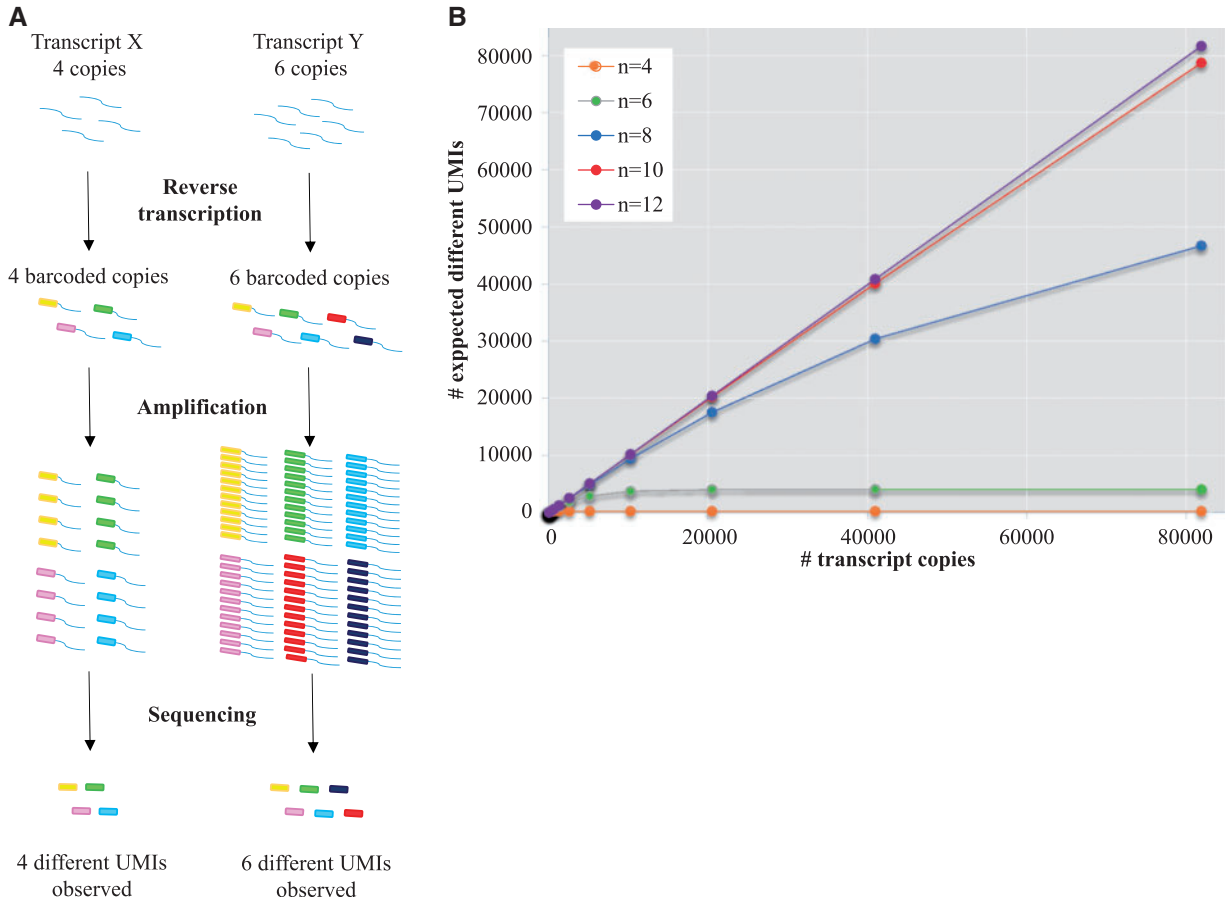

Figure 1. (**A**) In the hypothetical simple situation here represented, transcript X and Y are expressed with four and six copies, respectively. Consistently, four and six different UMIs (filled rectangles) are attached to the copies of transcript X and Y, respectively. Both transcripts undergo amplification (although with different efficiency, higher for gene Y) and are sequenced. After sequencing and read mapping, four different UMI copies of transcript X and six copies of transcript Y are detected. The figure is adapted from Islam *et al.* [22]. (**B**) Graph reporting on the *x*-axis the number of expressed transcripts copies and on *y*-axis the expected number of different UMIs observed, for different UMI lengths, based on the probabilistic relationship derived from [33]. The number of different UMIs is lower than the number of transcript copies for short UMIs and abundant transcripts as explained in section 'Transcripts quantification', Equations (1) –(3).

In principle, UMI-based protocols remove biases related to amplification and sequencing depth, as multiple reads associated with the same UMI and originated from the same transcript copy are collapsed into a unique count. However, this is only true if all libraries are sequenced at a sufficient depth, so that each uniquely tagged molecule is observed at least once. If not, some UMI-tagged cDNA molecules could be lost [34].

Finally, UMI-based protocols are generally cheaper and reliable; however, they are affected by coverage biases, as only one end of the transcript (the one containing the barcode) will be sequenced after fragmentation [21]. Therefore, UMI-based protocols are not good for monitoring splicing variants.

Bioinformatic analyses of scRNA-seq data

Like bulk RNA-seq, also scRNA-seq experiments generate FASTQ files, which contain thousands to millions of reads composed of RNA sequences and, eventually, add-on sequences (e.g. UMIs). With respect to bulk RNA-seq however, the typical scRNA-seq analysis workflow involves an additional cell quality control (QC) step and the analysis of the quantitative standards (Figure 2).

QC and reads alignment

The first step of scRNA-seq data analysis workflow is the preprocessing of the sequencing reads. The tools developed for QC and read alignment of bulk RNA-seq data may be used also for scRNA-seq [35–36]. Widely used preprocessing tools include Cutadapt [37], Trimmomatic [38], FASTQC [39] and Kraken [40]. More recently, also tools specifically developed for scRNA-seq reads QC have been proposed, such as sinQC [41] and Scater [42]. Most popular aligners include STAR [43], GSNAP [44], Tophat2 [45], HISAT [46] and the pseudo-aligner Kallisto [47]. For a benchmarking evaluation of alignment methods, please refer to Engström *et al.* [48] and Baruzzo *et al.* [49]. In case UMIs are used, it is necessary to trim the barcode sequence before performing read mapping; this can be performed for example with UMI-tools [50] or Je [51]. After read mapping, a second phase of QC is suggested. Suitable tools include packages like RseQC [52], SAMTOOLS [53] or Picard tools (https://broadinstitute.github.io/picard) or even one of the previously mentioned QC packages, which include pre- and post-mapping QC functions.

A specific step of the scRNA-seq data analysis workflow is the detection and the filtering of low-quality cells, which consists in excluding from further analyses dead cells, cells that have been broken during the capturing or cell doublets, i.e. pairs of cells sticking together and erroneously co-isolated. To this purpose, direct inspection through imaging approaches is often used before sequencing. In addition, some bioinformatics approaches have been developed, to be applied after sequencing and read mapping. SCell [54], for example, estimates the

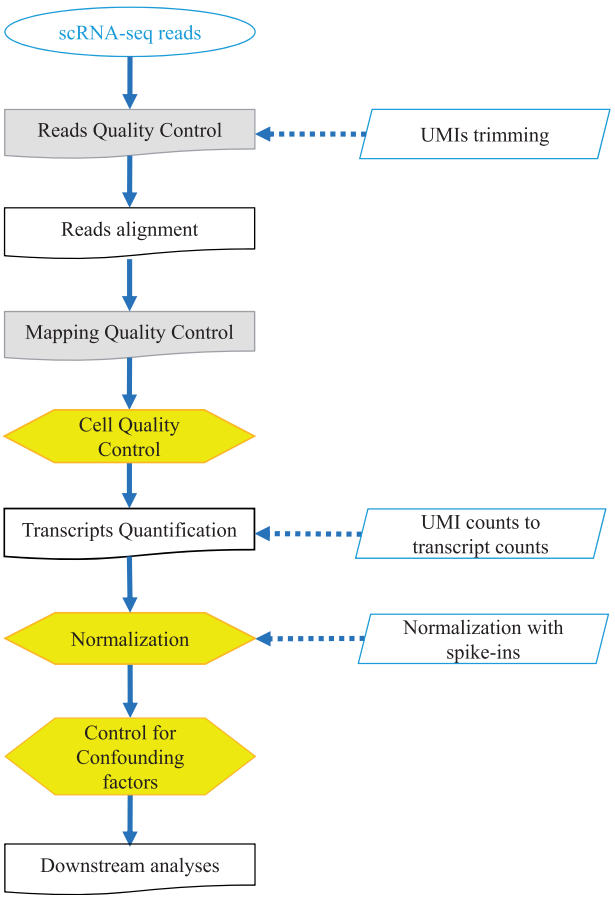

Figure 2. Schematic view of scRNA-seq data analysis workflow. Hexagonal boxes represent pivotal steps of scRNA-seq analysis, discussed in this review. Boxes linked by dashed arrows represent optional steps, depending on the scRNA-seq experimental protocol used.

number of genes expressed at background levels in a given sample and filters out cells whose background fraction is significantly larger than average. Low-quality cells can also be identified by analysing the proportion of reads mapped on spike-ins. A high ratio of reads mapping on spike-ins may indicate that the cell has broken during the capturing and that endogenous mRNA has been lost. Alternatively, in protocols where spike-ins are not used, low-quality cells can be detected by analysing the expression profiles with Cellity pipeline [55], which performs a model-based filtering of low-quality cells using a support vector machine trained on single-cell control samples.

Transcripts quantification

After mapping and quality checking, reads are ready to be summarized to generate expression levels. This could be done in a standard way by summing the reads mapping on each gene or using tools developed for bulk RNA-seq, such as HTSeq count [56], FeatureCounts [57], or maxcounts [58]. When instead UMIs are adopted, the number of UMI sequences has to be converted to transcript numbers. To this purpose, Grün *et al.* [59] introduced a statistical correction (*alias* 'collision probability'). This correction is useful especially when short UMIs ($n < 10$) are used, and the relationship between the number of total number of UMIs and the number of observed UMIs is non-linear and goes to saturation (Figure 1B).

Given m_i copies of transcript i and $k_{o,\,i}$ different UMIs observed, assuming a total of $K = 4^n$ (with n = UMI length) different UMI sequences are used and $K > m_i$ and sequencing depth sufficiently high, the probability of not observing UMIs ($K - k_{o,\,i}$) for transcript i is equal to:

$$\frac{(K - k_{o,i})}{K} \cong \left(1 - \frac{1}{K}\right)^{m_i}. \quad (1)$$

Thus, m_i can be estimated as:

$$m_i \cong \frac{\ln\left(\frac{K - k_{o,i}}{K}\right)}{\ln\left(1 - \frac{1}{K}\right)} \cong \frac{\ln\left(1 - \frac{k_{o,i}}{K}\right)}{\ln\left(1 - \frac{1}{K}\right)}, \quad (2)$$

which can be approximated as:

$$m_i \cong -K \cdot \ln\left(1 - \frac{k_{o,i}}{K}\right). \quad (3)$$

However, this method does not take into account for polymerase replication errors during PCR or errors in base calling during the sequencing, which may result in an overestimation of $k_{o,i}$, especially for longer UMI sequences. To address this issue, Islam *et al.* [22] proposed that UMIs with <1% of the average counts at the genomic locus should not be considered informative, whereas Bose *et al.* [59] proposed to merge all UMIs within a hamming distance of two or less. Alternatively, more recently Smith *et al.* [50] implemented three different methods to identify the number of unique molecules at a given locus: the cluster method [59], the 'adjacency method' and the 'directional adjacency method'.

Normalization of scRNA-seq count data

In single-cell RNA-seq experiments, each cell may be represented by a single biological system, a complex system that may be involved in several biological processes, including differentiation, cellular reprogramming and disease transformations, which imply transitions of the cells through distinct states [60]. Unbiased investigation of these cell states is challenging because of several factors, including the variability introduced by the cells' processing steps and by the sequencing. The systematic effects of technical sources of variation may be corrected in part through a process called normalization.

In bulk RNA-seq experiments, counts are scaled—either using median normalization or global scale factors—so that there would be, on average, no fold-difference in expression between cells for the majority of the genes [61, 62]. However, global scale factors have been shown to be not suitable for scRNA-seq, as they assume that the relationship between read counts and sequencing depth is common across genes and that they do not take into account for the high frequency of dropout events of scRNA-seq data [30]. Also normalizing for transcript length, e.g. FPKM normalization, may be problematic with current scRNA-seq protocols, as some of them are biased towards one end of the transcript (Table 2). Despite this conceptual unsuitability for scRNA-seq data, bulk normalization methods are still widely used.

Normalization with and without spike-ins

Normalization methods that exploit spike-in information are based on the idea that differences between the observed and the expected number of mapping spike-ins can be described in

terms of technical variation, as spike-ins are exposed to most of the experimental steps but are free of biological variation [30]. Thus, the proportions of mapping spike-ins at different expression levels are used to convert the number of endogenous mapping reads in the number of transcripts. This is the underlying idea of methods like SAMstrt [61] and GRM normalization [63]. In particular, GRM normalization fits a gamma regression model to the log-normalized spike-ins counts and derives the log-concentration of endogenous genes using the estimated parameters. Alternatively, other methods [64, 65] use spike-ins to calculate scaling factors and apply them to endogenous genes to obtain normalized expression estimates.

Other normalization approaches specific for scRNA-seq, which do not consider spike-ins, have been proposed. SCnorm [66] estimates for every gene the dependence of transcript expression on sequencing depth using quantile regression, by grouping genes with similar dependence. Then, it uses a second quantile regression to estimate scaling factors within each group and adjusts for sequencing depth. Lun *et al.* [67] proposed Scran, based on a deconvolution approach, to normalize on summed expression values from pools of cells. Briefly, summation of counts is performed across pools of cells to generate pool-based size factors, thus reducing the number of zeroes. All size factors from cell pools are then used to create a system of linear equations that is used for deconvolving the pool-based size factors, to infer the size factors for the individual cells.

Batch effects correction

One of the major contributors to the nuisance of scRNA-seq data is batch effects, systematic differences in gene expression levels, which arise from variability across experimental batches. Different experimental batches introduce a new source of variation in the data, which is consistent across transcripts and that can be confused with the biological signal of interest (e.g. different condition, different subpopulation) and prevents from appropriately modelling biological variation and group-specific changes in gene expression [68, 69]. A common strategy to assess the presence of batch effects is to use dimensionality reduction strategies (see section 'Data visualization'), such as principal component analysis (PCA), followed by diagnostic plots to see if groups of cells of the same batch tend to cluster together and separately from the other cells. Existing approaches for correcting for batch effects use empirical Bayes frameworks, such as ComBat [70], which is also robust to outliers for small sample sizes. More recently, Lun and Marioni [71] proposed the 'summation' approach, which implies summating the counts from all cells in each batch and use them instead of single-cell counts for downstream analyses (e.g. differential expression analysis). The underlying assumption of this method is that count sums are theoretically independent, as the batch effect is sampled independently for each batch, so the sums can be treated as replicate samples for each biological group.

Controlling for unwanted sources of variation

So far, we have reported methods that are used to correct for systematic errors, which consistently affect sequencing data. Data variability is also because of other factors, which are often referred to as 'confounding factors' (Figure 2). Factors such as the intrinsic transcriptional noise, because of stochastic and bursty transcription, and the extrinsic noise, because of, for example, the cell cycle or the differentiation state of the cell, can be considered 'confounding' as they prevent the biological signal of interest from being uncovered [35]. These sources of variation are not widespread throughout the transcriptome landscape [22] but involve only certain transcripts species; therefore, their effects do not have to be removed, like in normalization, but controlled or modelled. In bulk experiments, these genes are present as well and act with the same behaviour, but as the gene expression measurements are averaged, they do not significantly affect the downstream analyses as in scRNA-seq.

To identify genes that have higher variability with respect to that expected from technical sources (called 'High Variable Genes' or HVGs), Brennecke *et al.* [64] modelled the relationship between gene expression and the squared coefficient of variation (SCV), which reflects the variability in gene expression in relation to the mean expression level. They considered as HVGs those genes whose coefficient of variation significantly exceeded the 50% (i.e. SCV $>$ 0.25). Grün *et al.* [33] worked with a combination of UMIs and spike-ins to estimate both the technical and biological component of the variance. The technical component had the variance equal to the mean (as in a Poisson distribution), whereas the biological component had constant coefficient of variation (CV) across different expression levels. Moreover, Grün *et al.* [33] observed that the CV intensity depended on the total number of sequenced RNAs per sample, with a linear correlation of 0.91. This model has been shown to yield precise noise estimates consistent with single-molecule fluorescence in situ hybridization [9]. Later, Kim *et al.* [72] implemented the variance decomposition method, to subtract the technical variance calculated using spike-ins and estimate gene-specific biological variability used, in turn, to identify HVGs. Alternatively, Vallejos *et al.* [73] proposed the Bayesian Analysis of Single-Cell Sequencing Data (BASiCS). The authors modelled spike-ins and endogenous genes expression as a two Poisson-gamma hierarchical model with shared parameters, and estimated gene-specific posterior probabilities to identify both lowly and highly variable genes.

A second group of methods has been developed to account for the noise caused by genes with oscillatory behaviour, for example genes involved in the cell cycle. Indeed, during the cell cycle, a cell increases in size, replicates its DNA and splits into daughter cells [74]. Different cells, even being of the same type, could be at different time points of the cell cycle, thus having different gene expression profiles. Buettner *et al.* [66] proposed a Gaussian process latent variable model, which estimates the covariance matrix associated with the cell cycle oscillations using the expression profiles of 892 annotated cell-cycle genes. They aimed at estimating the underlying biological variability, first by adjusting for technical variation using spike-ins parameters estimates and then by adjusting for variation derived from the oscillatory genes. Alternatively, ccRemover [74] has been proposed to remove the principal components (PCs) affected by the cell cycle, and Oscope [75] has been proposed to detect genes with oscillatory behaviour, without a priori knowledge on which are the oscillatory genes, by combining a paired-sine model and K-medoids clustering.

Data visualization

The high number of genes and cells measured in a typical scRNA-seq experiment introduces data visualization challenges related to data projection into lower dimensions and clustering of data into putative cell subpopulations [76]. Indeed, when analysing cells in such high-dimensional gene space, it becomes more difficult to distinguish the difference between them, also because of the high level of noise. A widely used solution for single-cell data visualization is represented by dimensionality

reduction. Here, we discuss dimensionality reduction in relation to data visualization only, and not to downstream analyses (e.g. clustering, trajectory analysis), which is beyond the objectives of this review.

Dimensionality reduction implies the projection of cells data in a lower-dimensional space, and it is fundamental for cell quality checking, data inspection before and after normalization, outlier detection and confounding effect identification [35]. However, different methods give different results, given that the projection implies the loss of some information and, on the other hand, the prioritization of specific characteristics of the data [76].

The most widespread dimensionality reduction approach is PCA [77]. PCA uses a linear transformation to convert a set of observations into a set of values of linearly uncorrelated variables called PCs. These components explain decreasing amounts of variation in the data, so usually the first two or three PCs are used for visualization [78]. Variations of PCA have been recently proposed to take into account also for zero-inflation [79].

Another commonly applied method in scRNA-seq data visualization is the t-distributed stochastic neighbour embedding (t-SNE) [80], which combines dimensionality reduction with random walks on the nearest-neighbour network to map the data into a lower-dimensional space while preserving local distances between cells [76, 78]. Differently from PCA, t-SNE is a non-linear and stochastic algorithm, which means that applying t-SNE to the same data set multiple times will produce different embeddings, which are sensitive also to the choice of a 'perplexity' parameter, which reflects the number of neighbours used to build the nearest-neighbour network [76]. Thus, it is suggested to run the algorithm multiple times to determine the appropriate perplexity value for a particular data set and improve the stability of results, and to use it only for visualization purpose and not as a dimensionality reduction method [76].

A further dimensionality reduction approach is the diffusion maps (DM), implemented for example in R package Destiny [81] DM is a non-linear projection method, which assumes a smooth nature of the data, where the distance between cells reflects the transition probability based on several paths of random walks between the cells. The algorithm implemented in Destiny includes also the imputation of dropouts.

Design of scRNA-seq experiments

When designing an scRNA-seq experiment, several variables should be taken into account, as the number of cells to be sequenced, the isolation technique, the experimental protocol, the inclusion of quantitative standards and the required sequencing depth.

Basically, all of these aspects strongly depend on the biological question and, not less relevant, from the economic availability. A schematic view of three possible scRNA-seq experimental designs based on three different types of biological questions is reported in Figure 3.

In general, if the aim of the study is the identification of subpopulations in a large number of cells (i.e. thousands of cells), less 'sensitive' protocols—in terms of mRNA capture efficiency and transcript coverage—and lower sequencing depth may be used. Indeed, when analysing large cell populations with the aim of identifying different cell types, droplet-based protocols should be preferred to perform the sequencing of thousands of cells at lower depth [24]. Indeed, even if characterized by low capture efficiency, droplet-based protocols allow cell barcoding and pooling of the samples to generate a single sequencing library, thus reducing the cost of the experiment when large numbers of cells are analysed. As an example of experimental costs, to analyse 1000 cells with 150 paired-end reads and a depth between 250 000 and 500 000 reads per cell, the sequencing with a NextSeq500 High Output run has a total cost of ∼5000$, as by using droplet-based protocols, the library preparation cost is reduced down to ∼0.1$per cell [27]. Of course, if the economical availability is not an issue, full-length protocols and/or high-sequencing depth can be used with a high number of cells, consistently increasing of the cost of the experiment.

Indeed, when the purpose of the study is analysing transcript isoforms or characterizing transcription factors, fewer cells are needed but greater sequencing depth and more sensitive protocols, like template switching-based methods, are required to obtain whole-transcript coverage, thus increasing the cost of the experiment [30]. As an example, to analyse 100 cells with 150 paired-end reads at a depth between 3 and 5 M reads per cell, the price of the NextSeq500 High Output sequencing run (around 5000$) must be summed to the library preparation cost. If full-length protocols are used, one library for each cell is used. Thus, considering a library cost of ∼25–30$per cell depending on the kit used [27], the total cost for library preparation is ∼2500–3000$ and the total cost for the experiment is ∼8000$. A lower experimental cost is achievable by using in-house reagents instead of the commercial kits for library preparation, thus lowering the cost down to ∼3–5$ per cell [27, 83].

The use of quantitative standards in scRNA-seq experiments is highly recommended to improve the accuracy of the results [4, 33, 84]. When the aim of the study is to investigate differential gene expression or heterogeneity between subpopulation of cells, the use of UMI-based protocols has been demonstrated to increase the precision [17]. UMIs also reduce the technical variability, as they reduce amplification noise in the data, but, as previously said, they should not be used if the aim of the study is the analysis of splicing isoforms or SNP variants, as UMIs cause strong coverage bias towards the 5′ or 3′ end of the transcripts. Spike-in RNA molecules are useful instead for the QC of the cells and the normalization and to calculate the relation between cell sizes and RNA content (if relevant to the biological question), even if their quantity has to be accurately calibrated (see section 'Use of quantitative standards').

The total number of analysed cells, the use of quantitative standards and the amount of transcripts contained in the final sequencing library constrain the choice of the appropriate sequencing depth, which should be also carefully calibrated. Pollen *et al.* [85] state that 50 000 reads per cell are sufficient to detect the majority of genes that contribute to the overall population variance and that a sequencing depth between 5000 and 50 000 reads per cell is sufficient to detect cell subpopulations. However, at low depth, it is not possible to detect most of the lowly expressed genes [85]. Shalek *et al.* [86] state that 1 million reads per cell are sufficient to accurately estimate the mean and variance of gene expression [30]. Consistently, Tung *et al.* [87] suggest that using a sequencing depth of 1.5 million reads per cell is sufficient to be able to monitor also the lowly expressed genes. Rizzetto *et al.* [83] suggest a minimum read length of 100 bp paired-end with sequencing depth >250 000 reads per cell for accurate detection of gene expression or identification of cell subpopulations, and for minimizing the technical noise.

Another aspect that has to be carefully taken into account during sequencing is avoiding confounding batch effects. If multiple populations are being assessed, samples should be randomized even at sample preparation step and across

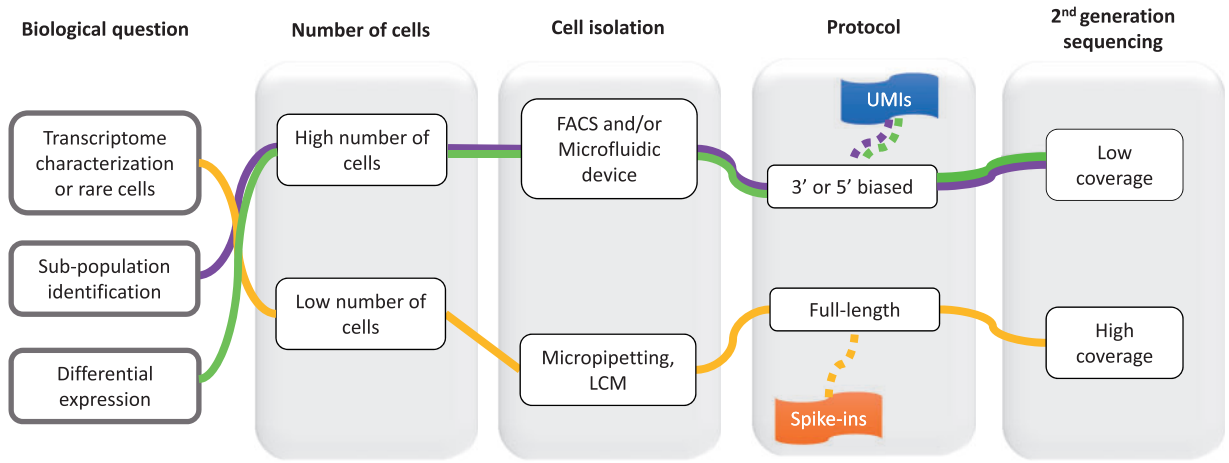

Figure 3. Workflow diagram of possible experimental designs of a scRNA-seq experiment. Dashed lines represent optional choices. Figure inspired by Cannoodt *et al.* [82].

Table 3. Main sources of bias in scRNA-seq experiments and solutions for limiting their impact

Source of bias	Type	Effect	Current solutions
RNA capture and RT efficiency	Technical	Stochastic zeroes	Spike-ins, statistical modelling
cDNA amplification	Technical	Loss of quantification accuracy	UMIs, statistical modelling
Batch effects	Technical	Introduce a signal different from the true biological signal	Statistical modelling
HVGs, transcriptional burst	Biological	Increase variance in the data	Statistical modelling
Cell-cycle stage, differentiation state, etc.	Biological	Confuse the true biological signal	Cell visualization, statistical modelling

multiple lanes within sequencing [30]. Unfortunately, when the number of cells is low, or time and budget constraints are limiting, the randomization is difficult to realize [30]. In such cases, it is necessary to correct for batch effects after sequencing, using specific statistical models (Table 3).

Conclusions

Single-cell RNA-sequencing is a promising technology, allowing the study of the transcriptome of single cells with unprecedented resolution and affordable costs. Anyway, single-cell data present high variability, deriving both from technical and biological sources (Table 3).

In the past few years, many single-cell isolation techniques, experimental protocols and bioinformatics tools have been developed, attracting remarkable attention to this field. In this review, we have discussed advantages and disadvantages of each method in relation to the scRNA-seq experimental design. In summary, from the technical perspective, no strategy has emerged to be the best one, so the choice should be done in relation to the underlying biological question. There is still way to go to overcome challenges deriving from cell isolation, mRNAs capturing and experimental protocols, but the use of quantitative standards and specific statistical methods surely helps in reducing or modelling the technical bias that characterizes scRNA-seq data (Table 3).

From the methodological perspective, even if several tools have been specifically developed for the normalization of scRNA-seq data, their performance has not been extensively benchmarked yet. Thus, many scRNA-seq analyses are still performed with methods originally developed for bulk RNA-seq, even if their adaptability to single-cell transcriptomics is not clear, as, for example, they do not take into account for dropout events [34]. Moreover, several tools developed for the analysis of scRNA-seq data address specific issues (e.g. effect of HVGs or of the cell cycle) but do not take into account other aspects, such as the percentage of transcripts with expression level of zero across different cells. The benchmarking of existing methods and the development of more sophisticated tools to improve current analysis strategies are strongly needed.

As no official guidelines for scRNA-seq experiments have been published yet, in this review, we aimed to highlight the strengths and weaknesses of existing experimental and bioinformatics approaches, suggesting possible scRNA-seq experimental designs based on different biological questions and giving estimates of the required budget. We hope this will help researchers planning a scRNA-seq experiment with greater awareness, to take full advantage of this powerful technology, which is increasingly being used by the scientific community.

Key Points

- scRNA-seq is a powerful tool to study individual cell transcriptomes on a large scale; however, the data generated by this technology are characterized by high levels of noise.
- Various experimental and computational methods for handling scRNA-seq data have been proposed; however, most of them are tailored to different experimental designs or biological questions.
- This review provides an overview of both scRNA-seq experimental and bioinformatic approaches, highlighting strengths and weaknesses of each approach.
- Possible experimental designs and analysis frameworks are proposed depending on specific biological questions and budget constraints, together with an evaluation of open challenges and future perspectives in the field.

Funding

This work was supported by the University of Padova, Department of Information Engineering, Project CPDR150320/15 ('Systems biology approach to single cell RNA sequencing') and Project PROACTIVE 2017 ('From Single-Cell to Multi-Cells Information Systems Analysis').

References

1. Dalerba P, Kalisky T, Sahoo D, *et al*. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011; **29**:1120–7.
2. Eberwine J, Sul J-Y, Bartfai T, *et al*. The promise of single-cell sequencing. *Nat Methods* 2014;**11**(1):25–7.
3. Patel AP, Tirosh I, Trombetta JJ, *et al*. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014;**344**(6190):1396–401.
4. Kolodziejczyk AA, Kim JK, Svensson V, *et al*. The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015;**58**(4):610–20.
5. Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 2014;**42**(14):8845–60.
6. Liang J, Cai W, Sun Z. Single-cell sequencing technologies: current and future. *J Genet Genomics* 2014;**41**(10):513–28.
7. Luni C, Giulitti S, Serena E, *et al*. High-efficiency cellular reprogramming with microfluidics. *Nat Methods* 2016;**13**(5):446–52.
8. Grün D, Van Oudenaarden A. Design and analysis of single-cell sequencing experiments. *Cell* 2015;**163**(4):799–810.
9. Macosko EZ, Basu A, Satija R, *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**(5):1202–14.
10. Habib N, Avraham-Davidi I, Basu A, *et al*. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017;**14**(10):955–8.
11. Tang F, Barbacioru C, Wang Y, *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**(5):377–82.
12. Sasagawa Y, Danno H, Takada H, *et al*. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *bioRxiv* 2017; doi:10.1101/159384.
13. Picelli S, Björklund ÅK, Faridani OR, *et al*. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 2013;**10**(11):1096–8.
14. Soumillon M, Cacchiarelli D, Semrau S, *et al*. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv* 2014.
15. Gierahn TM, Wadsworth MH, Hughes TK, *et al*. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* 2017;**14**(4):395–8.
16. Rosenberg AB, Roco C, *et al*. Scaling single cell transcriptomics through split pool barcoding. *bioRxiv* 2017; doi.10.1101/105163.
17. Hashimshony T, Senderovich N, Avital G, *et al*. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* 2016; **17**:77.
18. Islam S, Kjällquist U, Moliner A, *et al*. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 2011;**21**(7):1160–7.
19. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the last decade. *arXiv* 2017, arXiv:1710.05086.
20. Cao J, Packer JS, Ramani V, *et al*. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**:661–7.
21. Hebenstreit D. Methods, challenges and potentials of single cell RNA-seq. *Biology* 2012;**1**(3):658–67.
22. Islam S, Zeisel A, Joost S, *et al*. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;**11**(2):163–6.
23. Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**(5):1187–201.
24. Poulin J, Tasic B, Hjerling-Leffler J, *et al*. Disentangling neural cell diversity using single-cell transcriptomics. *Nat Neurosci* 2016;**19**(9):1131–41.
25. Picelli S. Single-cell RNA-sequencing: the future of genome biology is now. *RNA Biol* 2017;**14**(5):637–50.
26. Aird D, Ross MG, Chen WS, *et al*. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011;**12**(2):R18.
27. Ziegenhain C, Vieth B, Parekh S, *et al*. Comparative analysis of single-cell RNA-sequencing methods (Sup). *Mol Cell* 2017;**65**(4):631–43.e4.
28. Baker SC, Bauer SR, Beyer RP, *et al*. The external RNA controls consortium: a progress report. *Nat Methods* 2005;**2**(10):731–4.
29. Svensson V, Natarajan KN, Ly L-H, *et al*. Power analysis of single cell RNA-sequencing experiments. *Nat Methods* 2017;**14**(4):381–7.
30. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* 2016;**17**:63.
31. Paul L, *et al*. SIRVs: spike-in RNA variants as external isoform controls in RNA-sequencing. *bioRxiv* 2016.
32. Hardwick SA, Chen W, Wong T, *et al*. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Methods* 2016;**13**(9):792–8.
33. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014;**11**(6):637–40.
34. Vallejos CA, Risso D, Scialdone A, *et al*. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat Methods* 2017;**14**(6):565–71.
35. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015; **16**:133–45.
36. Poirion OB, Zhu X, Ching T, *et al*. Single-cell transcriptomics bioinformatics and computational challenges. *Front Genet* 2016;**7**:163.
37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**(1):10–2.
38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15):2114–20.
39. Andrews S. FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinformatics* 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
40. Davis MPA, van Dongen S, Abreu-Goodger C, *et al*. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* 2013;**63**(1):41–9.
41. Jiang P, Thomson JA, Stewart R. Quality control of single-cell RNA-seq by SinQC. *Bioinformatics* 2016;**32**(16):2514–6.
42. McCarthy DJ, Campbell KR, Lun ATL, *et al*. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 2017;**33**(8):1179–86.
43. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.

44. Wu TD, Reeder J, Lawrence M, *et al*. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol Biol* 2016;**1418**:283–334.
45. Kim D, Pertea G, Trapnell C, *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;**14**(4):R36.
46. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**(4):357–60.
47. Bray NL, Pimentel H, Melsted P, *et al*. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016; **34**:525–7.
48. Engström PG, Steijger T, Sipos B, *et al*. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 2013;**10**(12):1185–91.
49. Baruzzo G, Hayer KE, Kim EJ, *et al*. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods* 2017;**14**:135–7.
50. Smith TS, Heger A, Sudbery I. UMI-tools: modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* 2017;**27**:491–9.
51. Girardot C, Scholtalbers J, Sauer S, *et al*. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* 2016;**17**(1):419.
52. Benjamini Y, Speed TP. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 2012;**40**:e72.
53. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
54. Diaz A, Liu SJ, Sandoval C, *et al*. SCell: integrated analysis of single-cell RNA-seq data. *Bioinformatics* 2016;**32**(14):2219–20.
55. Ilicic T, Kim JK, Kolodziejczyk AA, *et al*. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* 2016; **17**:29.
56. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2014;**31**:0–5.
57. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**(7):923–30.
58. Finotello F, Lavezzo E, Bianco L, *et al*. Reducing bias in RNA sequencing data: a novel approach to compute counts. *BMC Bioinformatics* 2014;**15 (Suppl 1)**:S7.
59. Bose S, Wan Z, Carr A, *et al*. Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biol* 2015; **16**:120.
60. Armond JW, Saha K, Rana AA, *et al*. A stochastic model dissects cell states in biological transition processes. *Sci Rep* 2015;**4**(1):3692.
61. Katayama S, Töhönen V, Linnarsson S, *et al*. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 2013;**29**:2943–5.
62. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Brief Funct Genomics* 2015;**14**(2):130–42.
63. Ding B, Zheng L, Zhu Y, *et al*. Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 2015; **31**(13):2225–7.
64. Brennecke P, Anders S, Kim JK, *et al*. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 2013; **10**(11):1093–5.
65. Buettner F, Natarajan KN, Casale FP, *et al*. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 2015;**33**(2):155–60.
66. Bacher R, Chu LF, Leng N, *et al*. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* 2017;**14**(6):584–6.
67. Lun LAT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016; **17**:75.
68. Leek JT. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* 2014;**42**:e161.
69. Hicks SC, Teng M, Irizarry RA. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv* 2015.
70. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**(1):118–27.
71. Lun ATL, Marioni JC. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* 2017;**18**:451–64.
72. Kim JK, Kolodziejczyk AA, Illicic T, *et al*. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat Commun* 2015;**6**:8687.
73. Vallejos CA, Marioni JC, Richardson S, Morris Q. BASiCS: bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* 2015;**11**(6):e1004333.
74. Barron M, Li J. Identifying and removing the cell-cycle effect from single-cell RNA-sequencing data. *Sci Rep* 2016;**6**:33892.
75. Leng N, Chu L-F, Barry C, *et al*. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat Methods* 2015;**12**(10):947–50.
76. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. *Mol Aspects Med* 2017;**59**:114–22.
77. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *Philos Mag Ser* 1901;**6**(2):559–72.
78. Rostom R, Svensson V, Teichmann SA, *et al*. Computational approaches for interpreting scRNA-seq data. *FEBS Lett* 2017; **591**(15):2213–25.
79. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* 2015;**16**:241.
80. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
81. Angerer P, Haghverdi L, Büttner M, *et al*. Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 2016; **32**(8):1241–3.
82. Cannoodt R, Saelens W, Saeys Y. Computational methods for trajectory inference from single-cell transcriptomics. *Eur J Immunol* 2016;**46**(11):2496–506.
83. Rizzetto S, Eltahla AA, Lin P, *et al*. Impact of sequencing depth and read length on single cell RNA sequencing data: lessons from T cells. *bioRxiv* 2017.
84. Kivioja T, Vähärautio A, Karlsson K, *et al*. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 2011;**9**(1):72–4.
85. Pollen AA, Nowakowski TJ, Shuga J, *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* 2014;**32**:1053–8.
86. Shalek AK, Satija R, Shuga J, *et al*. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 2014; **510**(7505):363–9.
87. Tung P-Y, Blischak JD, Hsiao CJ, *et al*. Batch effects and the effective design of single-cell gene expression studies. *Nat Sci Rep* 2017;**7**: doi:10.1038/srep39921.